The CFC and the CSD

How has the CSD evolved and what can the CFC expect next?

Suzanna Ward – Head of Data and Community

CFC15: March 2024

CCDC
advancing structural science

The Cambridge Crystallographic Data Centre

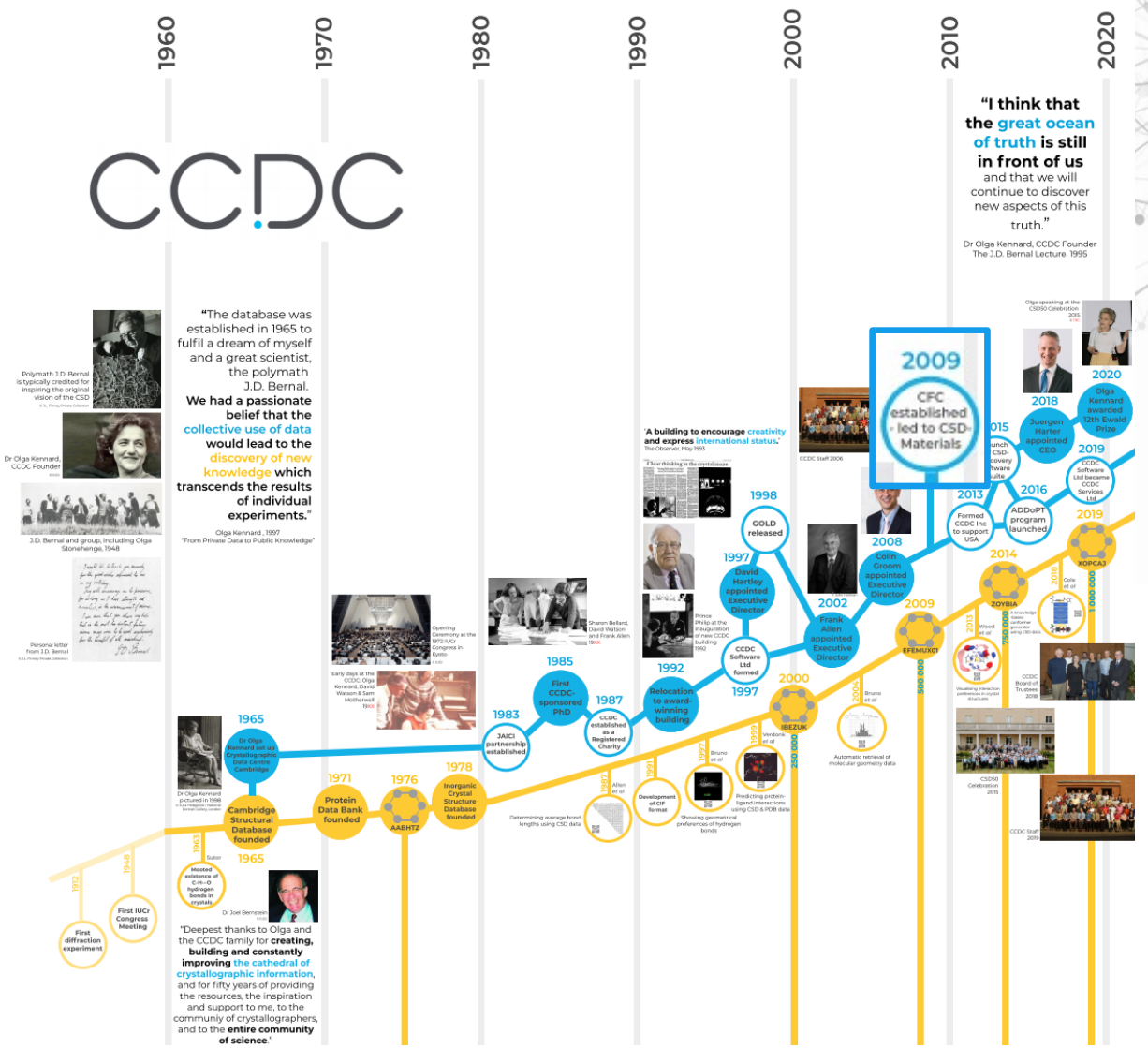
Polymath J.D. Bernal is typically credited for inspiring the original vision of the CSD

Dr Olga Kennard, CCDC Founder

J.D. Bernal and group, including Olga Stonehenge, 1948

"The database was established in 1965 to fulfil a dream of myself and a great scientist, the polymath J.D. Bernal. **We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments."**

Olga Kennard , 1997
"From Private Data to Public Knowledge"

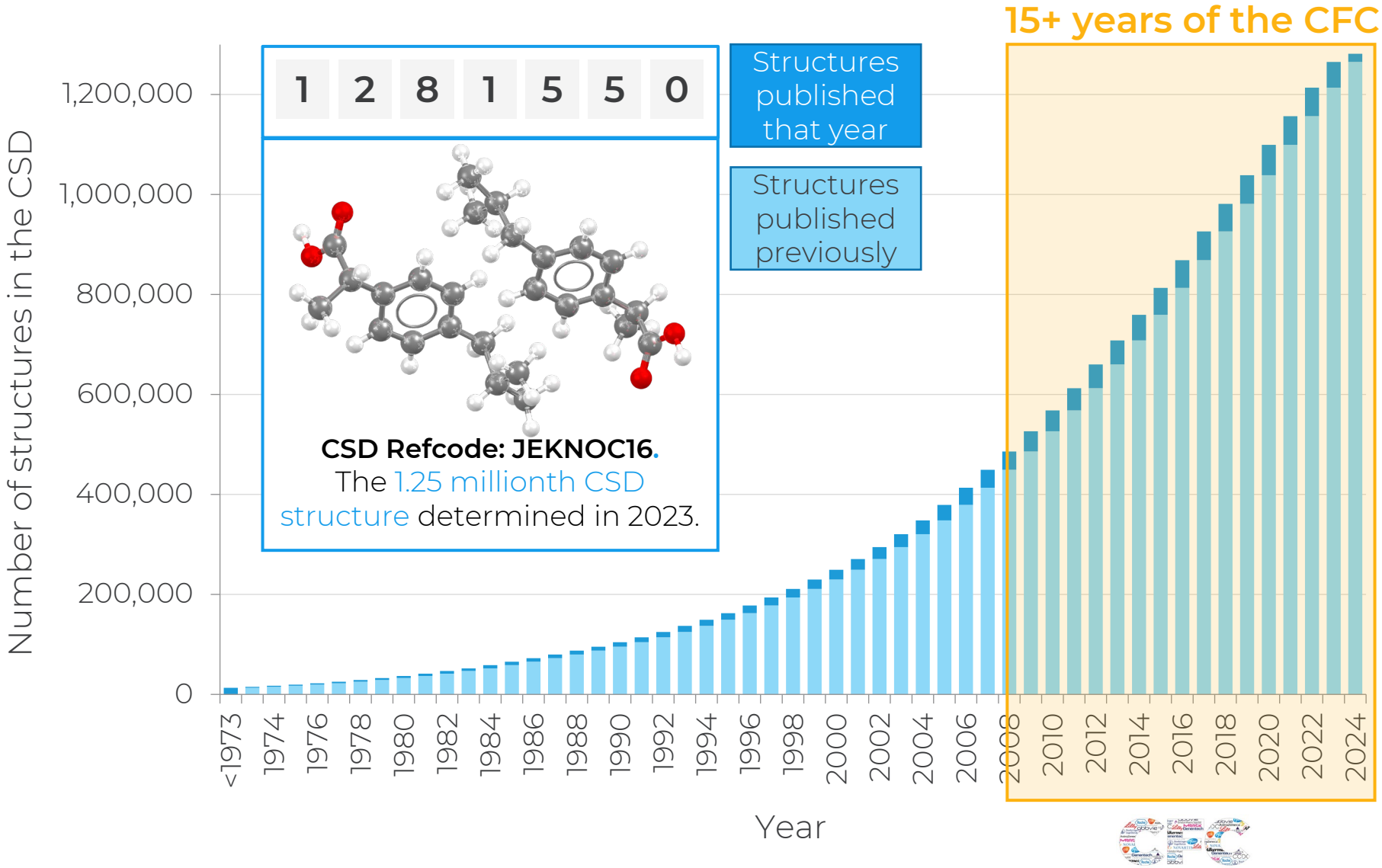


A wealth of information

- > 1 million structures
  - > 100M 3D coordinates
- > 28 million bond lengths
  - > 2M unique distributions
- > 40 million valence angles
  - > 3M unique distributions
- > 14 million torsion angles
  - > 800K unique distributions
- > 2 million rings
  - > 400K unique distributions
- > 2 million hydrogen bonds
  - >30 million Isostar contacts

Images and graphics created using the CSD Python API and Flourish

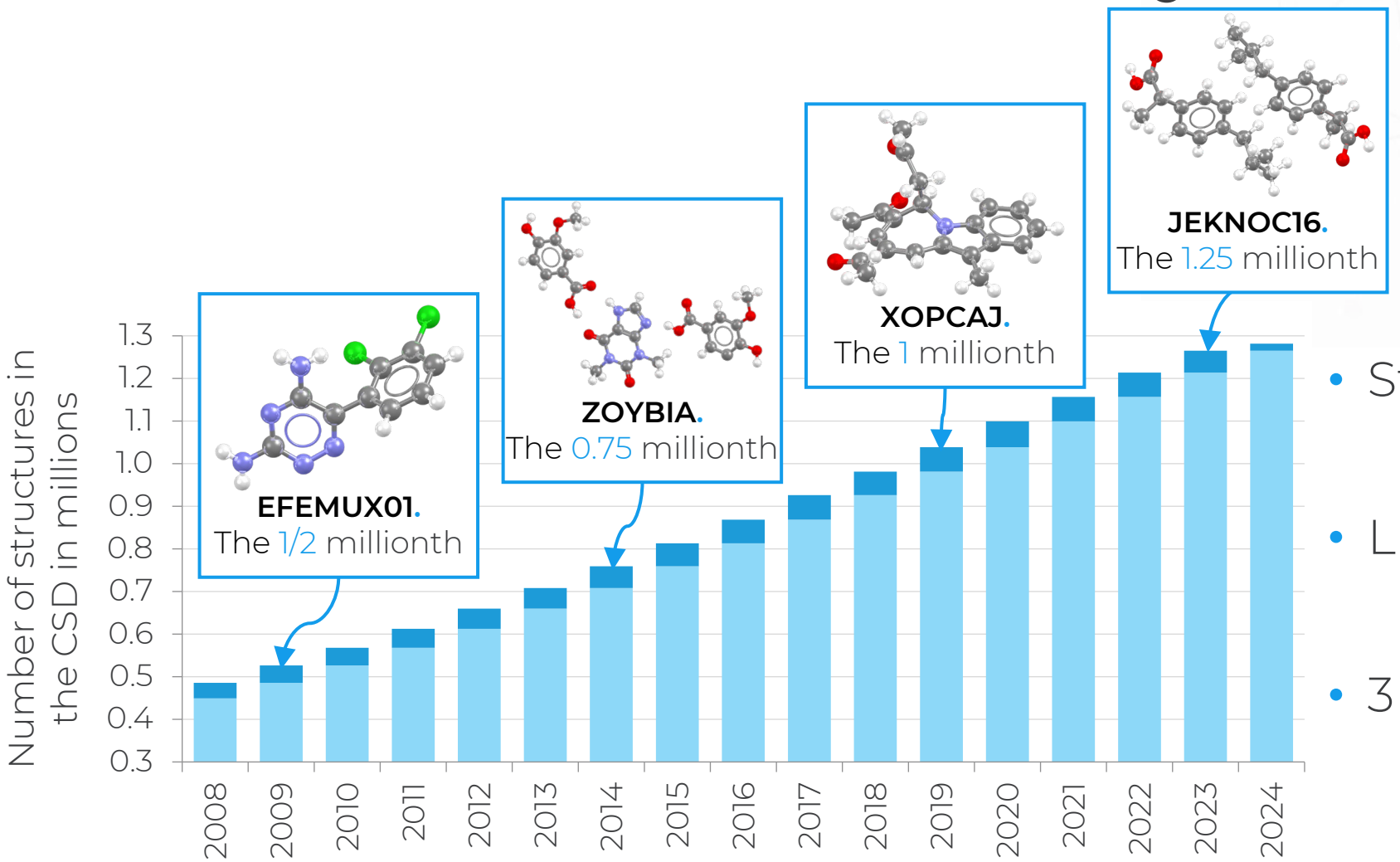
The Cambridge Structural Database



- Every published structure
  - Inc. ASAP & early view
  - *CSD Communications*
  - Patents
  - University repositories
  - Thesis
- Every entry enriched and annotated by experts
- Discoverability of data and knowledge
- Sustainable for over 58 years
- A trusted CoreTrustSeal repository

Certified as Trustworthy by CoreTrustSeal

The CSD and the CFC in 15 years

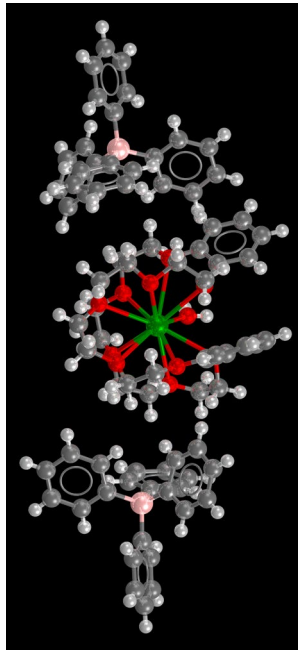


15 years in numbers

- Structures X 2.6
  - +>750K structures
- Literature sources X 1.6
  - +>730 literatures sources
- 3D atomic coords X 3.3
  - +>78M coordinates

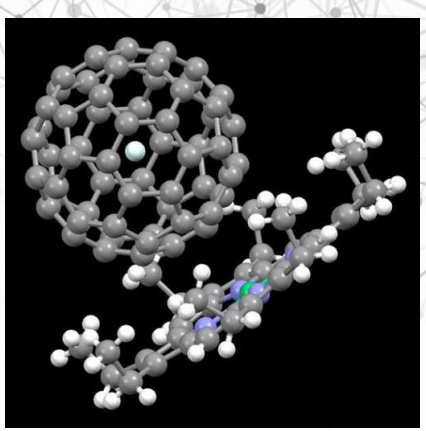
CCDC

New elements during the CFC years

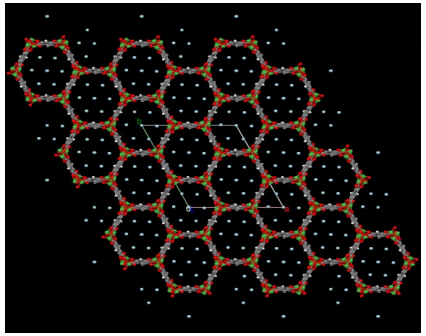


Radium added
2023
HIRCUK

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	2000s	CFC 2008-2024								
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	**	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
	*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
	**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	



Helium added 2013
YEMTUH



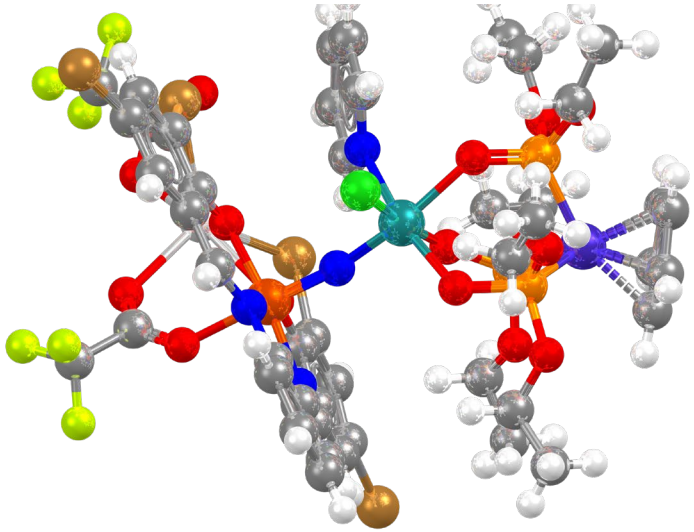
Neon added 2016
AQUCOG

Changing chemistry

KEDMEP

C_{46} H_{50} Ag Br_4 Cl Co F_6 Fe N_4 O_{15} P_3 Ru, 2(C H_2 Cl_2)

Most different elements in a structure (12)



Changing nature of substances

>42% Multiple component structures

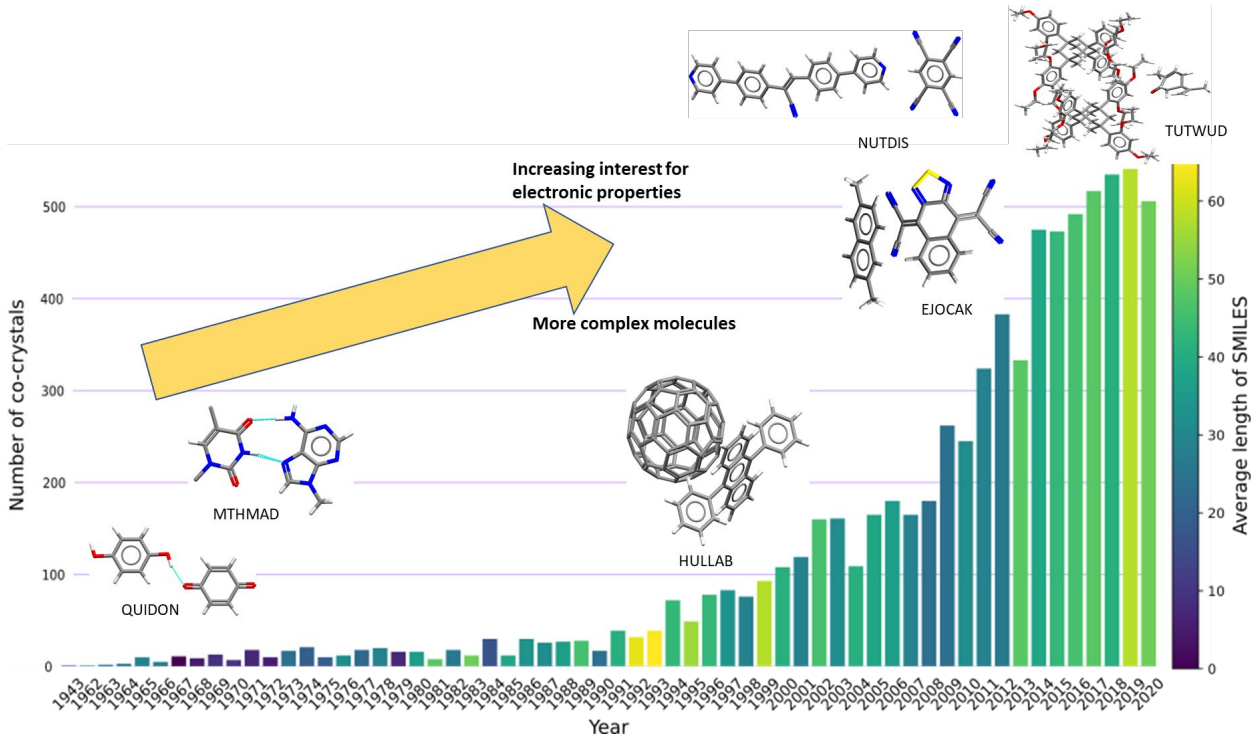
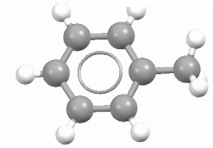
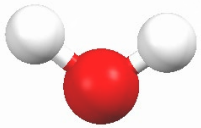


Chart showing the number of co-crystals in the CSD by year and colour coded by the average SMILES length. Image created by Katerina Vriza*, a CCDC sponsored PhD student.

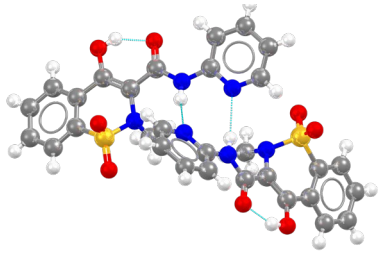


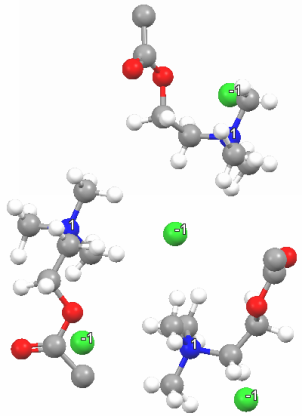
Solvates
>270K

Hydrates
>150K

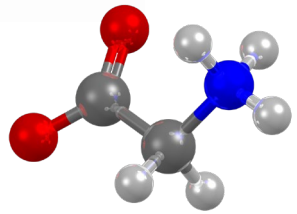


Polymorphs
>13K polymorphic families





Salts
>300K



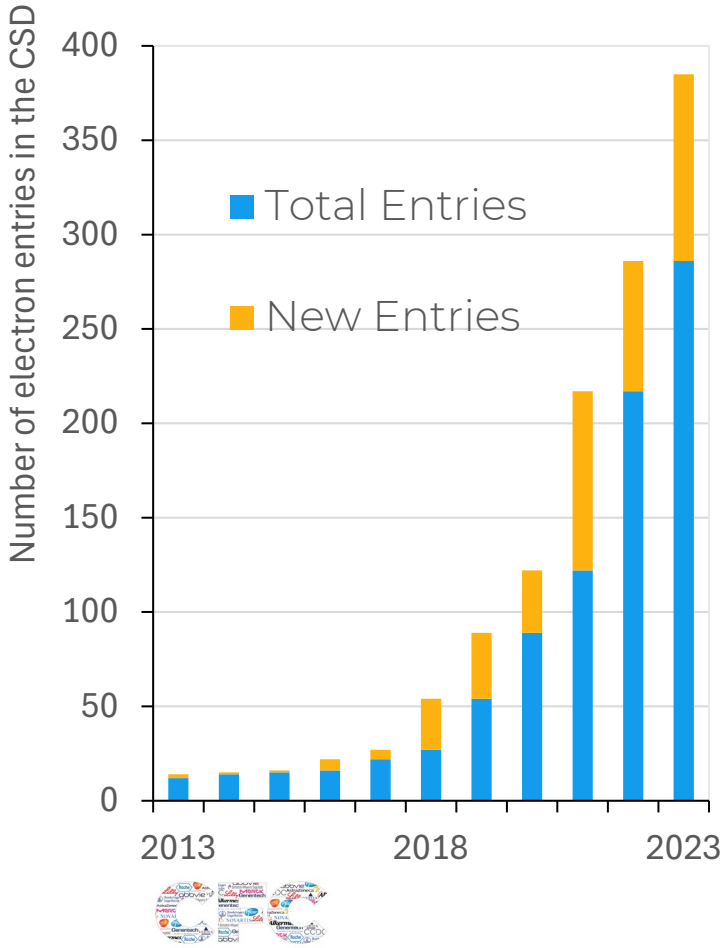
Refcode families
>1 Million

*Katerina Vriza, University of Liverpool, PhD on Data driven discovery of functional molecular co-crystal

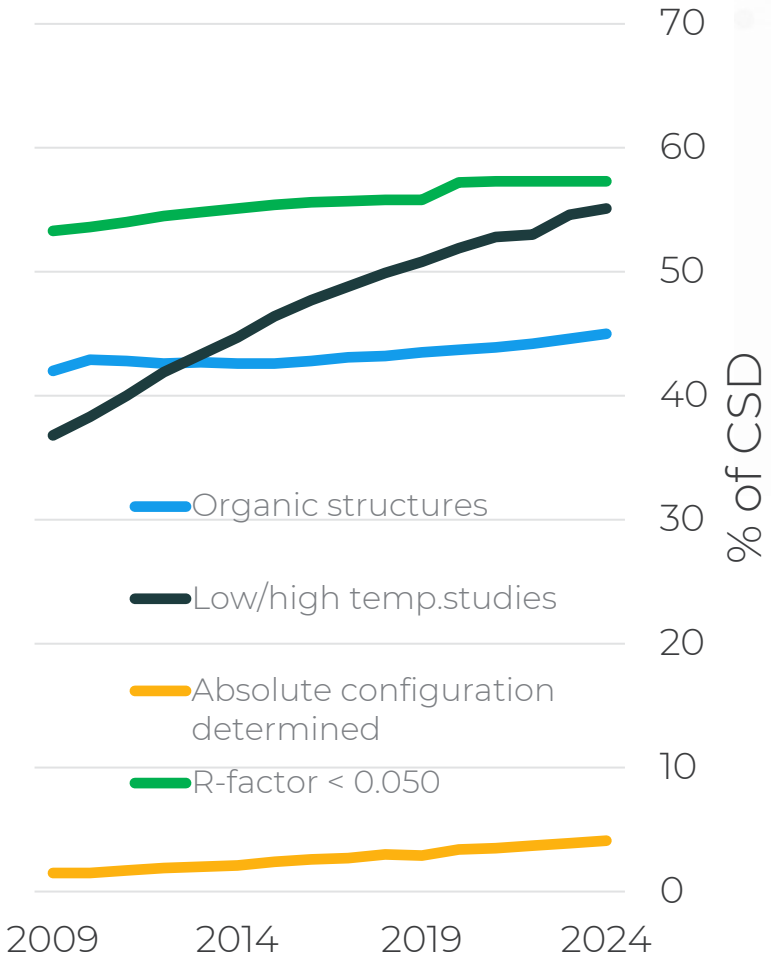
CCDC

Changing chemistry & crystallography

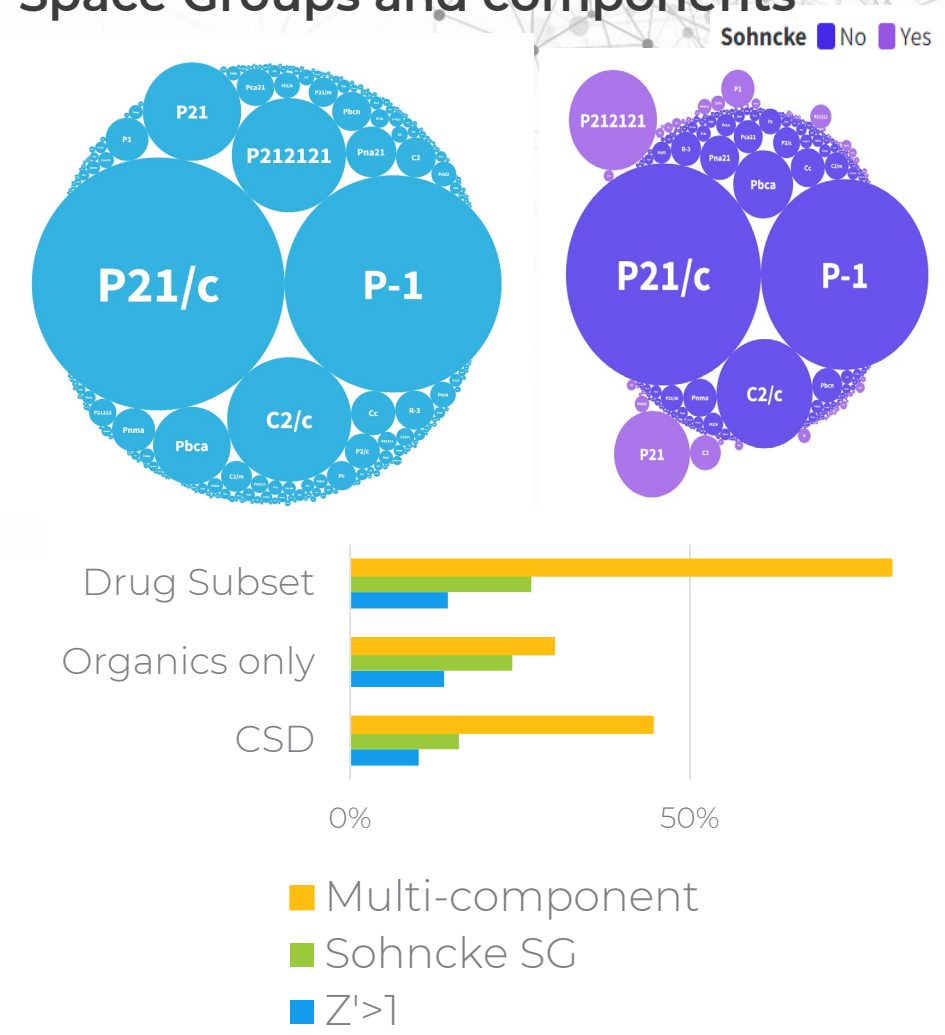
Growth of Electron Diffraction data



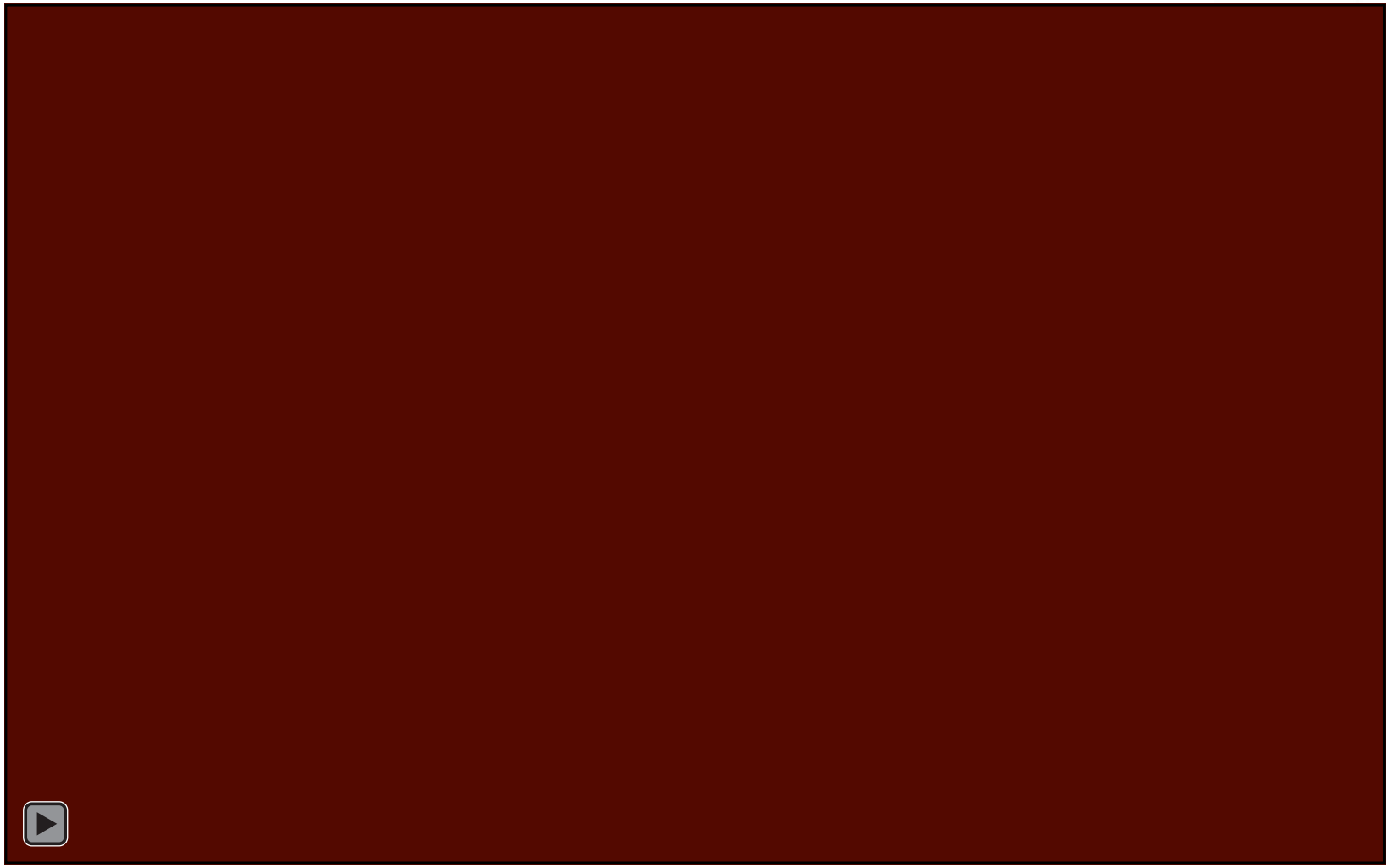
Changing metrics

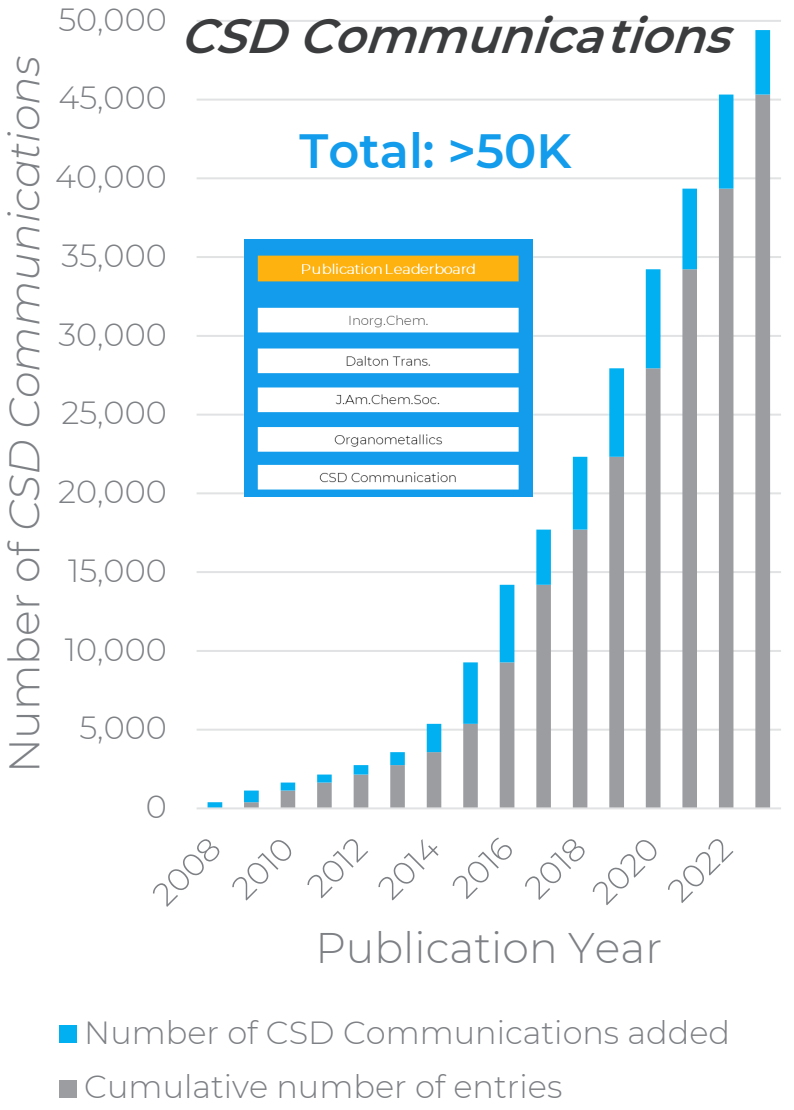


Space Groups and components



Changing people and places





Increasing data in the CSD

The CSD contains every published structure including:

- Inc. ASAP & early view
- University repositories
- Historical articles
- Patents
- Thesis
- *CSD Communications*

Paid summer placements

- Established >5 years ago
- 4 paid summer students a year
- Focussed on helping to improve wealth and value of data in the CSD

Last 15 years of the CFC

- >500 patents
- >350 thesis publications

Valuable data from CFC members

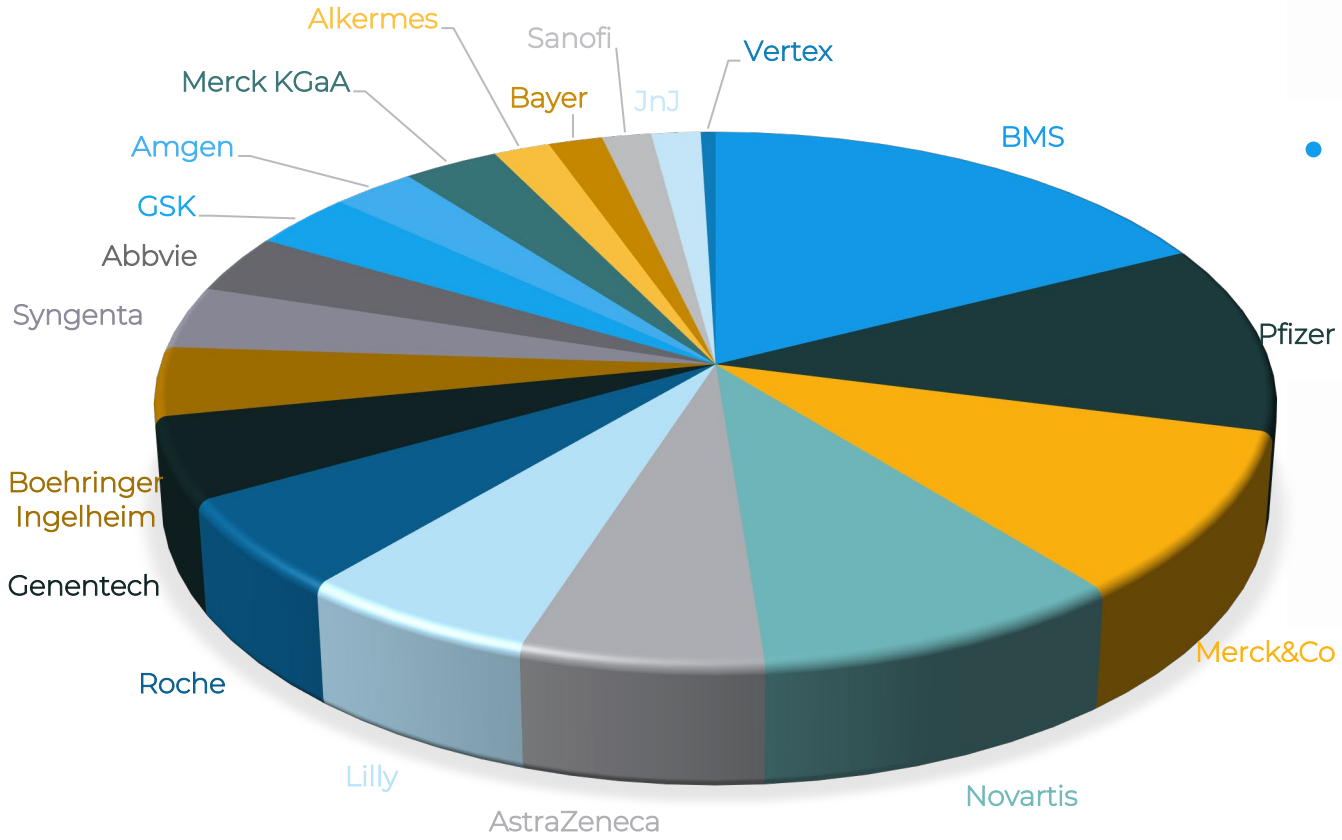
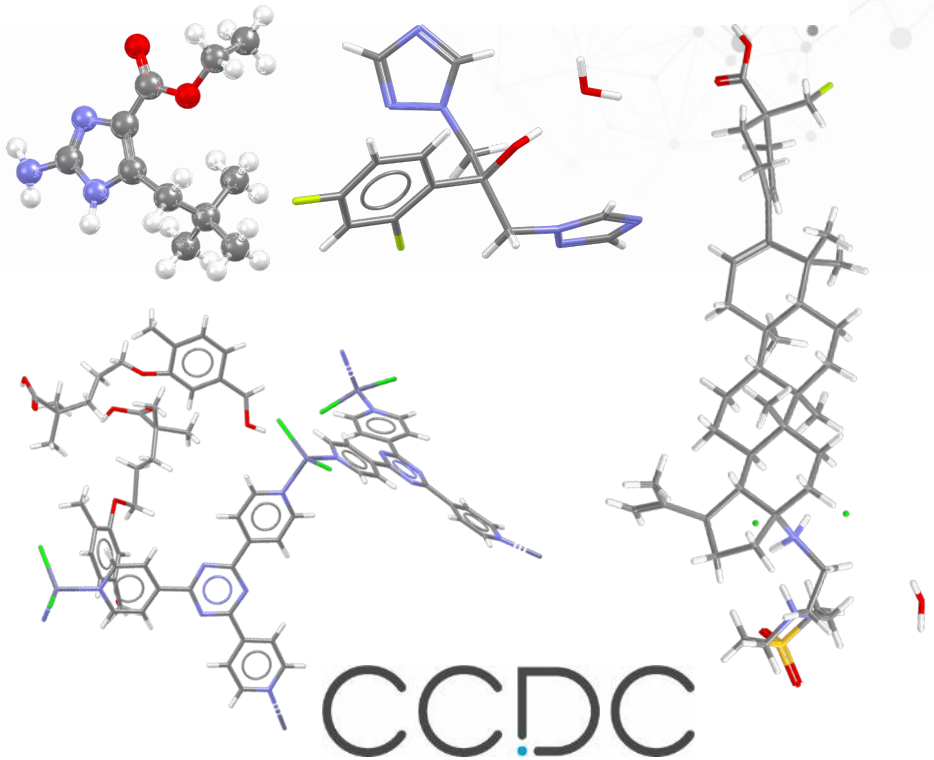


Chart created by searching for published structures in the CSD that have been deposited using a company email address or are associated to crystallographer details with a company email address

- **>1,240** structures
  - from past and present CFC members since 2008



CCDC

Small molecules, big impact

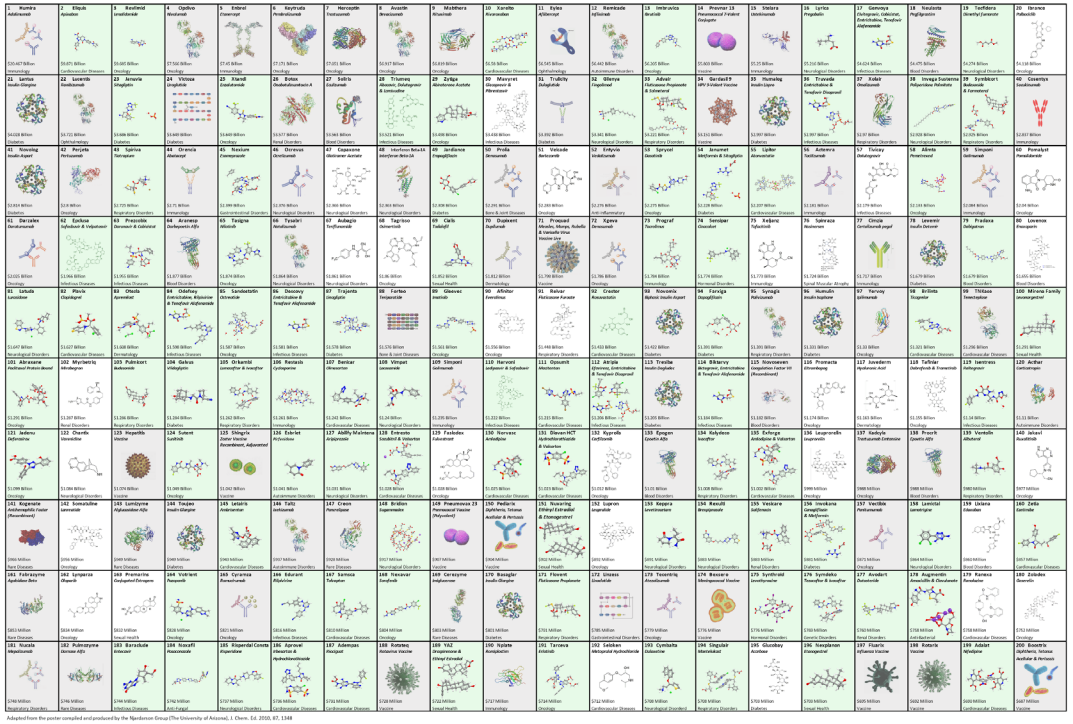
200 top drugs by retail sales in 2000

Small molecule structure in the CSD

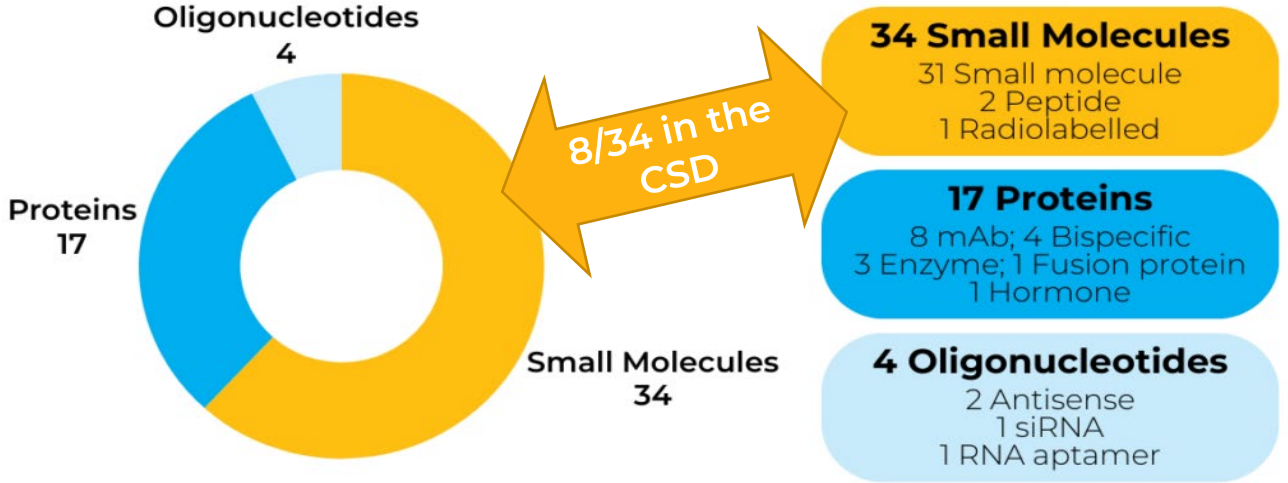
- Open-up new biological and therapeutic opportunities
- Compete against emergent modalities for rare diseases
- Are used to target RNA*

Small molecule structure CSD entry pending

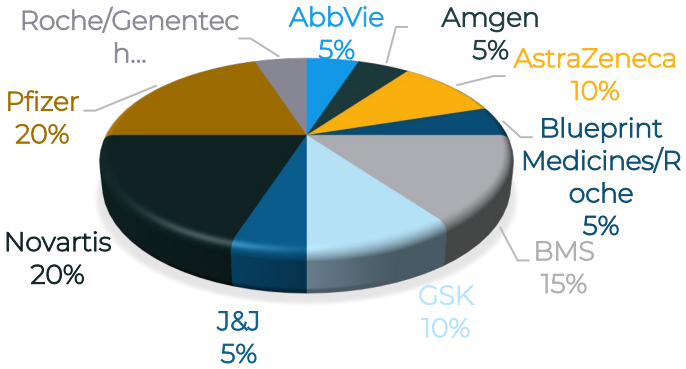
Biological structure in the PDB



FDA novel drug approvals 2023



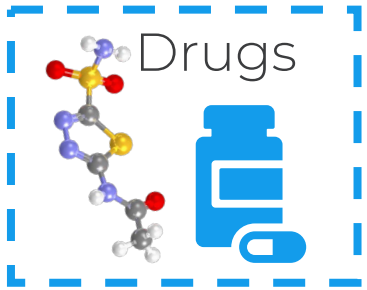
Missing FDA drugs 2020-23



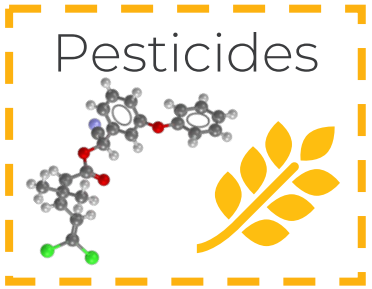
How can we help you to increase the amount of industrial data shared through the CSD?

Adapted from poster compiled and produced by the Njardarson Group (The University of Arizona)
Nature Reviews Drug Discovery **20**, 85-90 (2021)

The CSD Drug and Pesticide Subsets

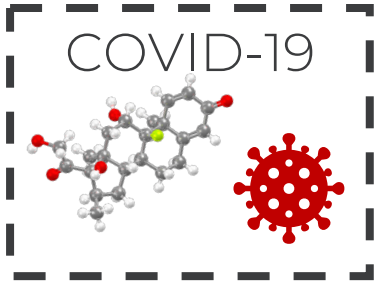


- Approved drugs in DRUGBANK
  - >15K in CSD Drug subset
    - +10K during 15 years of the CFC
  - >2K in Single-component subset

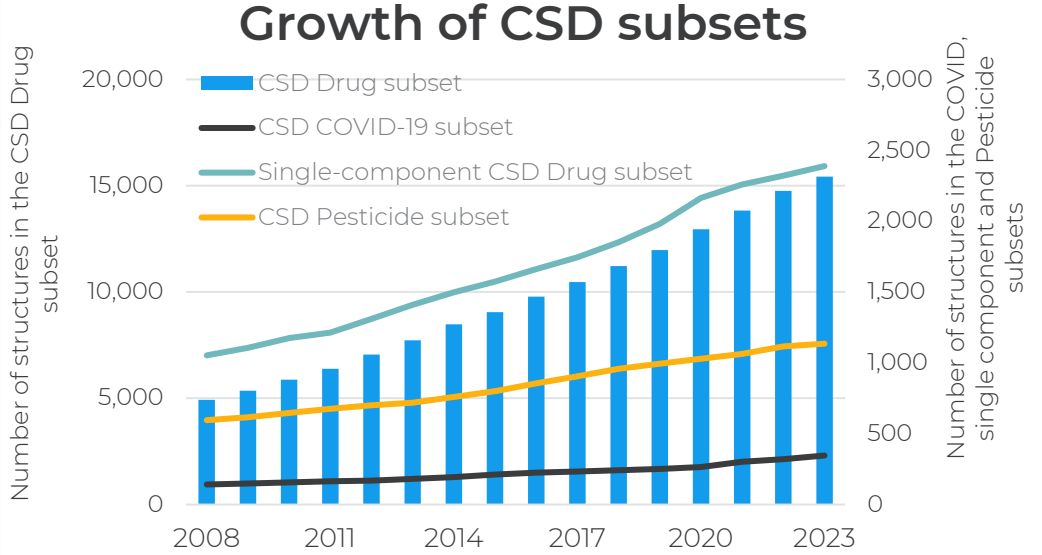


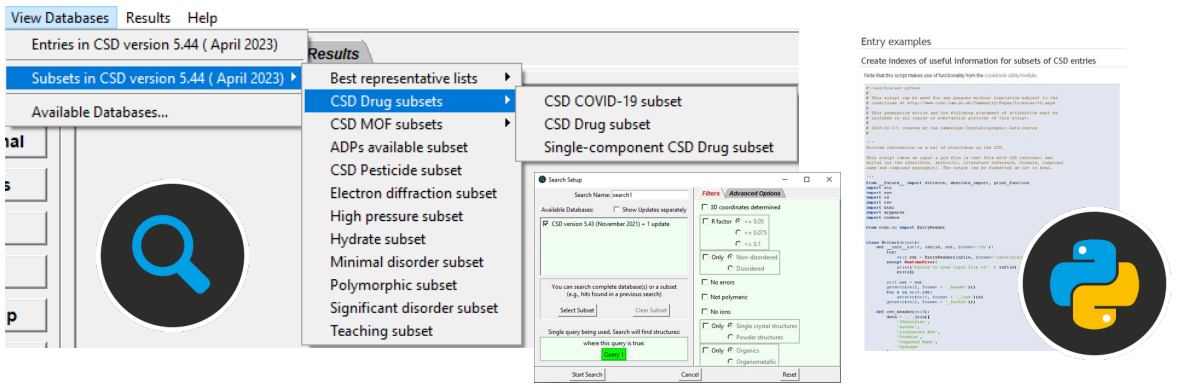
- In Pesticide Property DataBase
  - >1K in Pesticide subset

PPDB



- Identified as molecules of interest in fight against COVID-19
  - >340 in COVID subset





M.J. Bryant *et al.* *Journal of Pharmaceutical Sciences* 108 (2019) 1655-1662

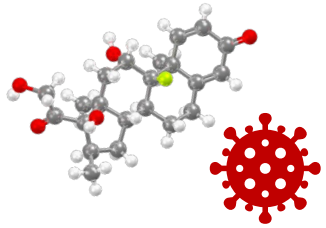
CCDC

CSD Subsets

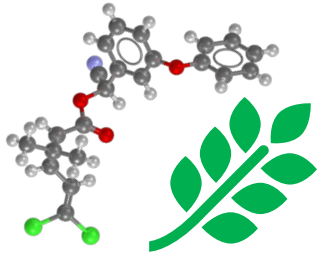
Groups of structures that may b more difficult to find in CSD fror searching alone

Added in 2020

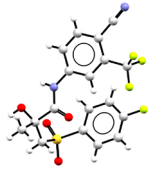
COVID-19



Pesticides

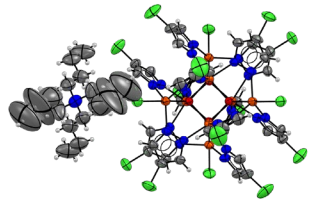


Drugs



Added in 2017

ADPs

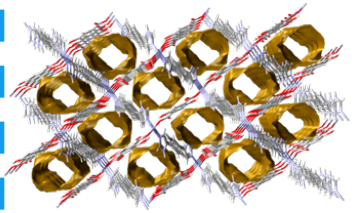


Added in 2010

Teaching

Added in 2016

MOFs

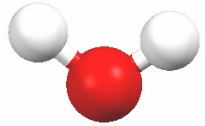


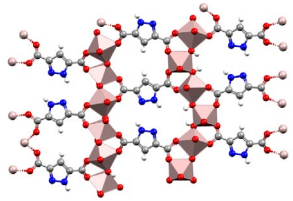
MOF Dimensionality

- 1D
- 2D
- 3D

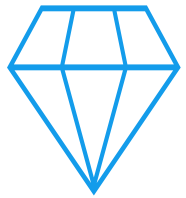
Added in 2022

Hydrates



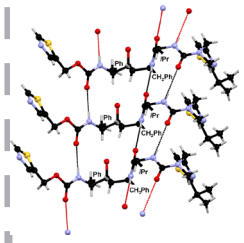


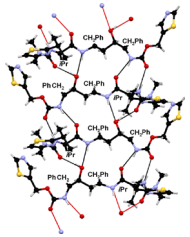
Electron Diffraction



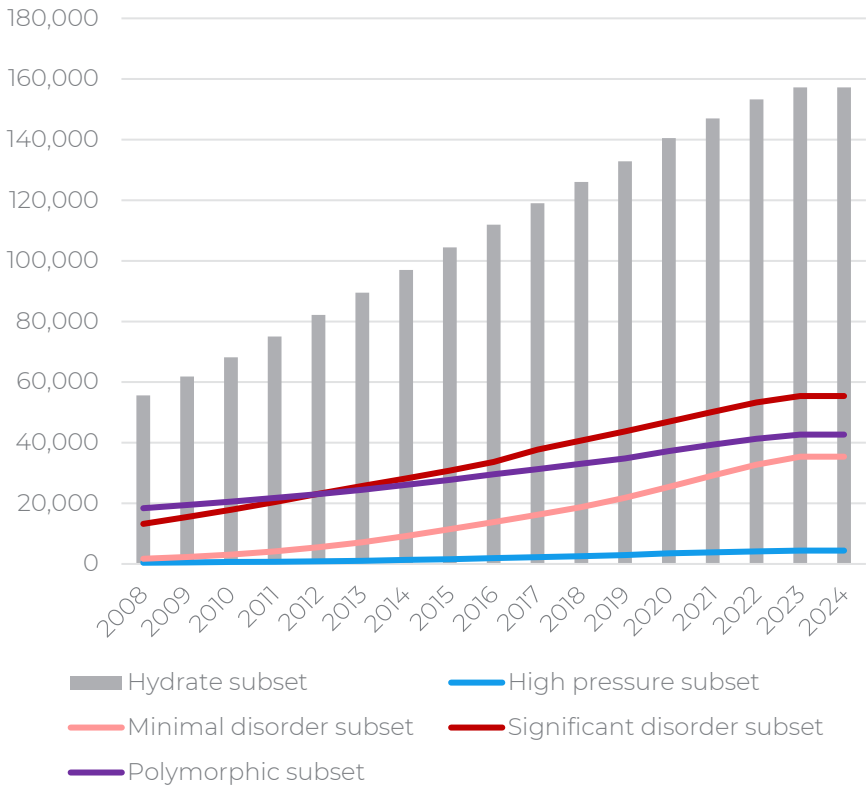
High Pressure

Polymorphs



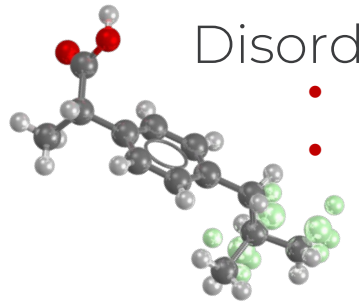


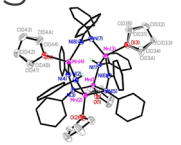
Growth of CSD subsets



Disorder

- Minimal
- Significant

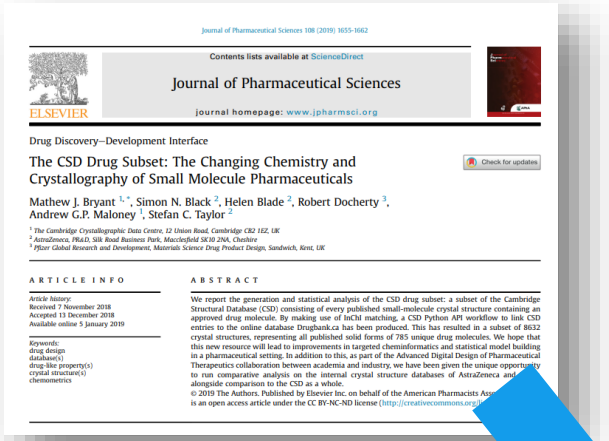




Added in 2022

CCDC

Your data collections


Journal of Pharmaceutical Sciences 108 (2019) 1655-1662

Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Drug Discovery–Development Interface

The CSD Drug Subset: The Changing Chemistry and Crystallography of Small Molecule Pharmaceuticals

Mathew J. Bryant [1,*], Simon N. Black [2], Helen Blade [2], Robert Docherty [3], Andrew G.P. Maloney [1], Stefan C. Taylor [2]

[1] The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK
[2] AstraZeneca, PR&D, Silk Road Business Park, Macclesfield SK10 2NA, Cheshire
[3] Pfizer Global Research and Development, Materials Science Drug Product Design, Sandwich, Kent, UK

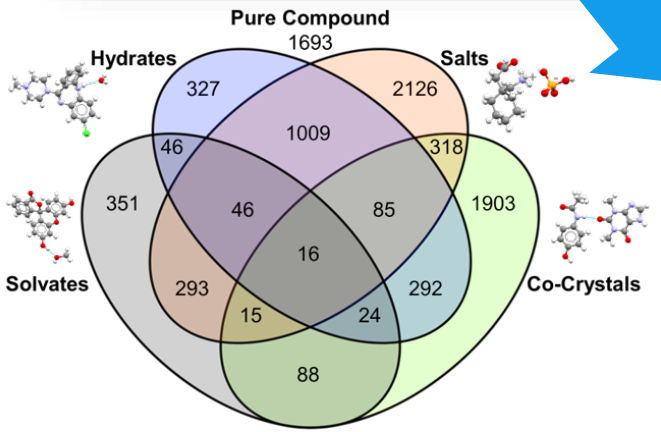
ARTICLE INFO

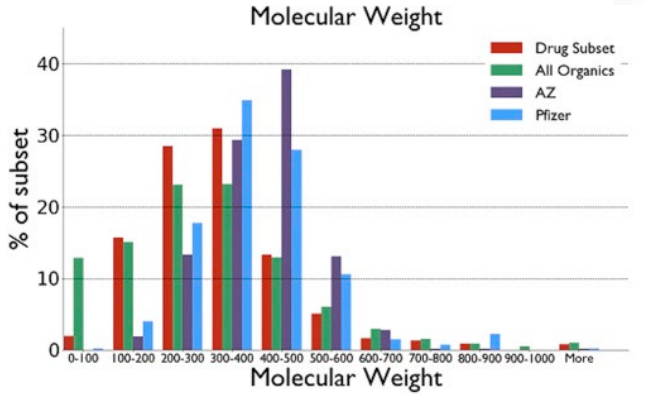
Article history:
Received 7 November 2018
Accepted 13 December 2018
Available online 5 January 2019

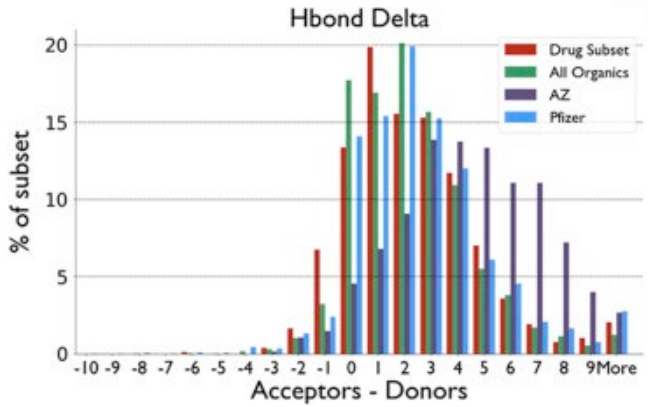
Keywords:
drug design
database(s)
drug-like property(s)
crystal structure(s)
chemometrics

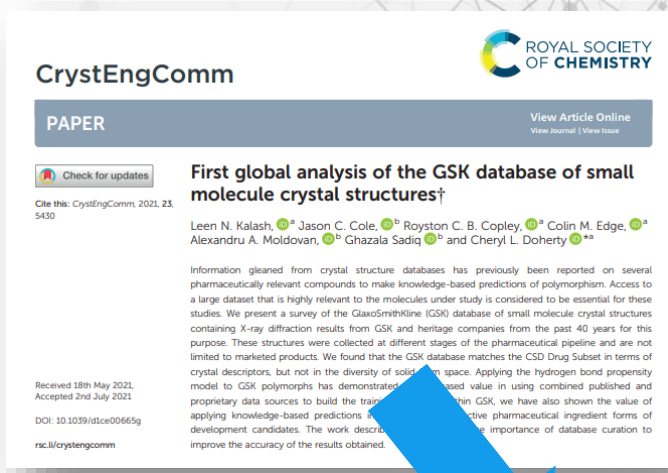
ABSTRACT

We report the generation and statistical analysis of the CSD drug subset: a subset of the Cambridge Structural Database (CSD) consisting of every published small-molecule crystal structure containing an approved drug molecule. By making use of InChI matching, a CSD Python API workflow to link CSD entries to the online database Drugbank.ca has been produced. This has resulted in a subset of 8632 crystal structures, representing all published solid forms of 785 unique drug molecules. We hope that this new resource will lead to improvements in targeted cheminformatics and statistical model building in a pharmaceutical setting. In addition to this, as part of the Advanced Digital Design of Pharmaceutical Therapeutics collaboration between academia and industry, we have been given the unique opportunity to run comparative analysis on the internal crystal structure databases of AstraZeneca and alongside comparison to the CSD as a whole.








CrystEngComm

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online

Check for updates

Cite this: *CrystEngComm*, 2021, 23, 5430

First global analysis of the GSK database of small molecule crystal structures†

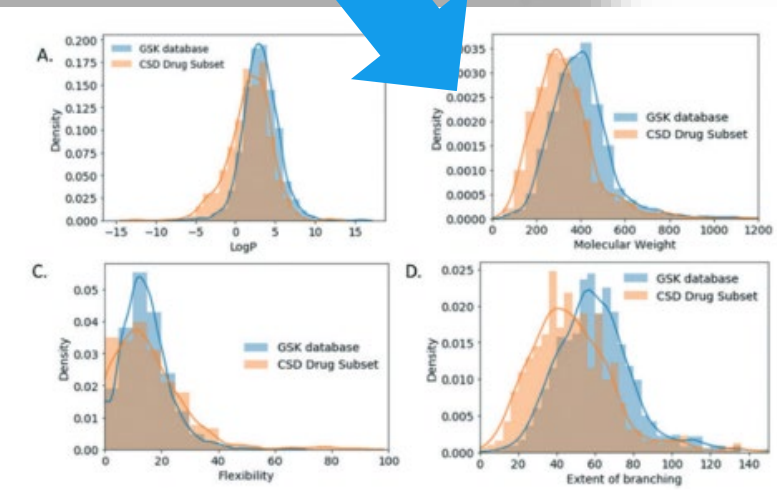
Leen N. Kalash, Jason C. Cole, Royston C. B. Copley, Colin M. Edge, Alexandru A. Moldovan, Ghazala Sadiq and Cheryl L. Doherty

Information gleaned from crystal structure databases has previously been reported on several pharmaceutically relevant compounds to make knowledge-based predictions of polymorphism. Access to a large dataset that is highly relevant to the molecules under study is considered to be essential for these studies. We present a survey of the GlaxoSmithKline (GSK) database of small molecule crystal structures containing X-ray diffraction results from GSK and heritage companies from the past 40 years for this purpose. These structures were collected at different stages of the pharmaceutical pipeline and are not limited to marketed products. We found that the GSK database matches the CSD Drug Subset in terms of crystal descriptors, but not in the diversity of solid [illegible] space. Applying the hydrogen bond propensity model to GSK polymorphs has demonstrated [illegible]ased value in using combined published and proprietary data sources to build the train[illegible]hin GSK, we have also shown the value of applying knowledge-based predictions [illegible]ctive pharmaceutical ingredient forms of development candidates. The work descri[illegible]e importance of database curation to improve the accuracy of the results obtained.

Received 18th May 2021,
Accepted 2nd July 2021

DOI: 10.1039/d1ce00665g

rsc.li/crystengcomm



M.J. Bryant *et al. Journal of Pharmaceutical Sciences* 108 (2019) 1655-1662
L.N.Kalash *et al. CrystEngComm* 23 (2021) 5430-5442

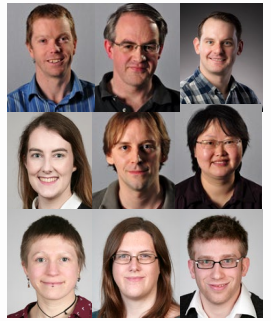
Proprietary data curation

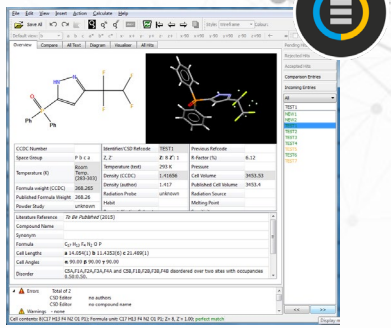
Data siloes and poor data quality can make finding, sharing and generating insights from data hard

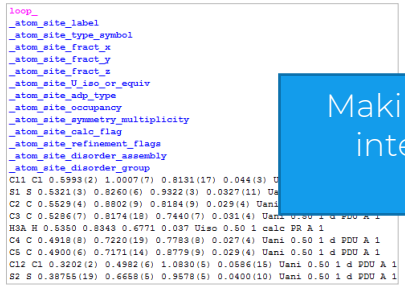
Curation, management & consultancy towards FAIR industrial data

Helping to store experimental and predicted results together

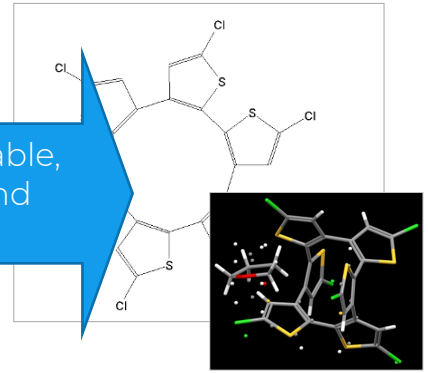
Ensuring structural data is accessible across the organisation







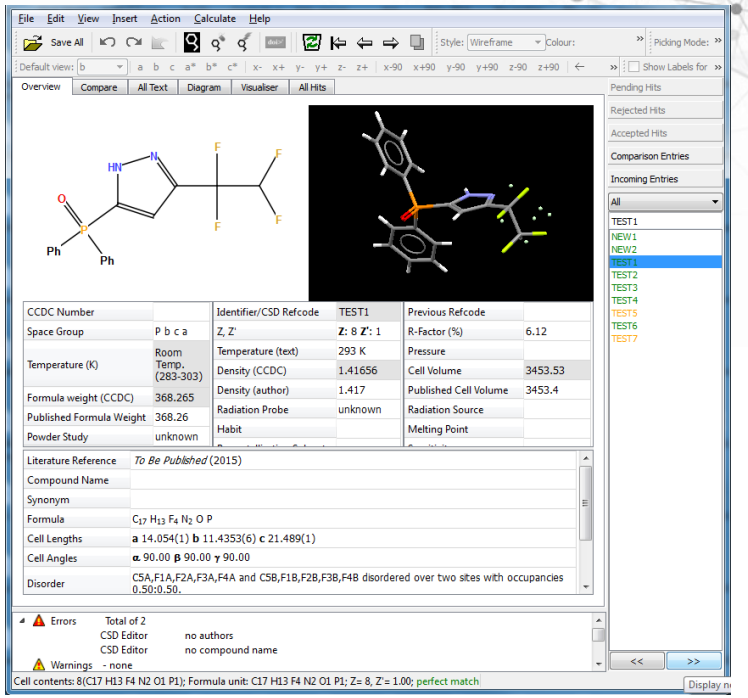
Making data findable, interoperable and reusable



CCDC

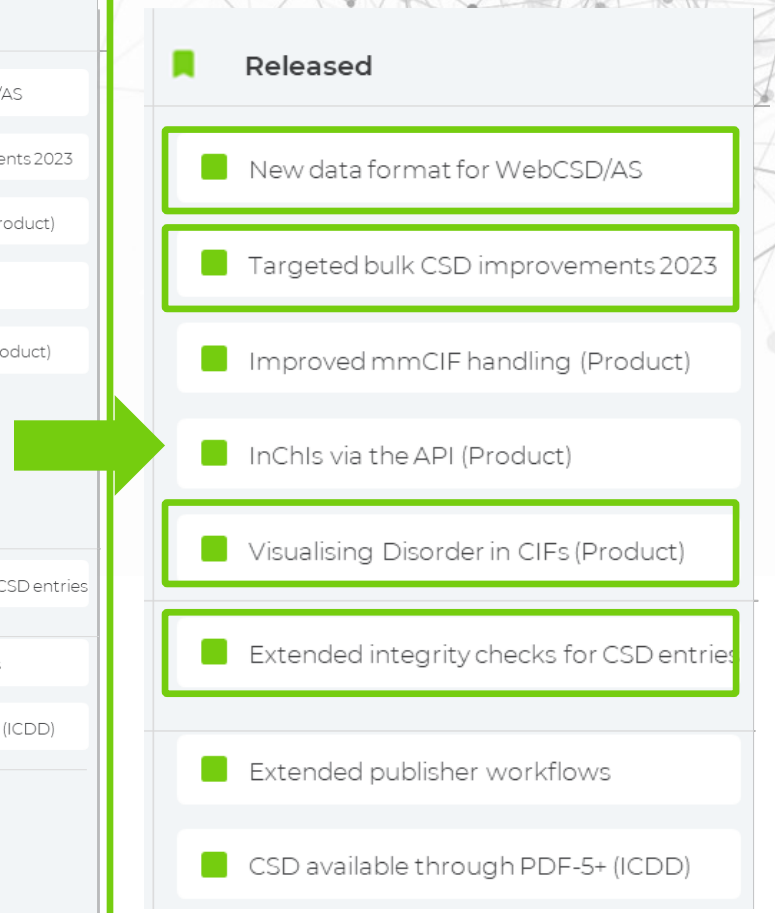
Supporting your data management

- We are available to you to create, enhance, update and maintain your databases of in-house structures or to help you in this process as a service
- Service work has included:
  - Building databases
  - Auditing, updating and improving existing databases
  - Providing bespoke database training
  - Writing custom API scripts
  - Enhancing existing databases
  - Creating databases of CSP/DFT structures
  - Building bespoke subsets of data in the CSD
  - Data quality checks and enhancements



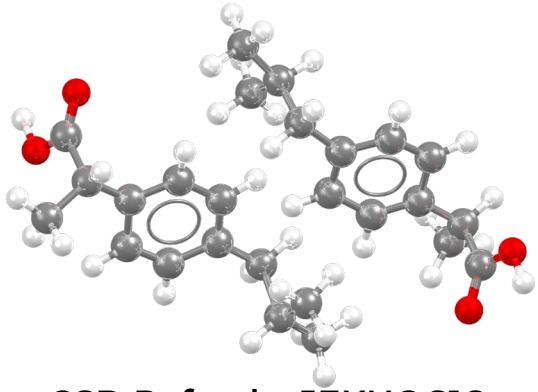
Where are we now and where next?

	Candidate	Planned	In progress	Released
Data	Capturing crystallisation conditions	New data fields available	Targeted bulk CSD improvements 2024	New data format for WebCSD/AS
	Data versioning	Open licenses for deposited CIFs	Calculated semi-conductor data	Targeted bulk CSD improvements 2023
	Evolving our knowledge bases	New data subsets		Improved mmCIF handling (Product)
	Improved metadata for PXRD, ED, etc	Visualising disorder for CSD entries		InChIs via the API (Product)
	Alternative ways to navigate/view entries	New data format for desktop release		Visualising Disorder in CIFs (Product)
	New models of data			
	Improved stereochemistry representations			
Data integrity	New integrity checks in deposition	Data integrity filters	Additional data integrity metadata	Extended integrity checks for CSD entries
Data partners		Improved and extended links	Further publisher workflows	Extended publisher workflows
		CCDC visualisation in more journals		CSD available through PDF-5+ (ICDD)
Data standards			Community CSP data standards	
			Schema for crystallisation conditions	
			CoreTrustSeal re-certified	

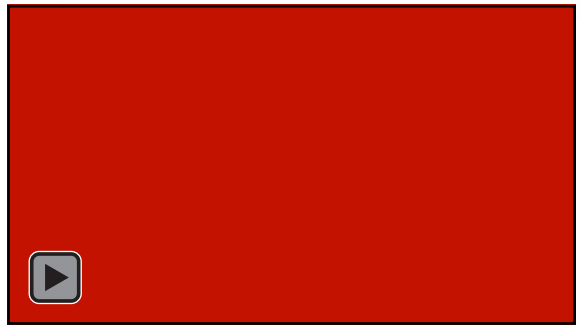


Data releases 2023-24

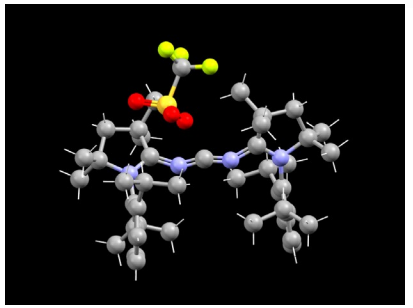
- **November 2023** – Full data update
  - **10,837 new structures** (11,411 new entries).
  - Taking the **total size** of the CSD to **1,254,560 structures** (1,284,316 entries).
  - Included new entries and **improvements** to over **190K existing entries**.
- **Coming soon!**
  - First quarterly data release for 2024
  - Data update for on-site WebCSD



CSD Refcode: JEKNOC16.
The 1.25 millionth CSD structure determined in 2023.



CSD Refcode: NIWYUR.



CSD Refcode: NITRUH.

Database and architecture evolution

Evolving the database at the heart of our software portfolio

- Distributed deployment via cloud based systems
- A unified platform and modern structure to facilitate searching
- Extended capacity to cater for increased data volumes
- Flexible and extendable to enable additional data types and features

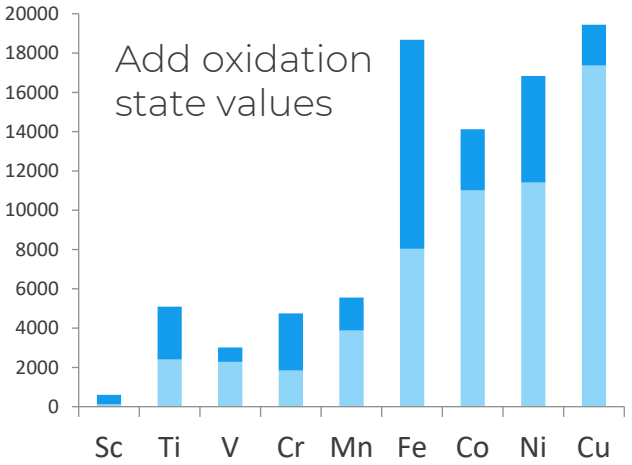
New database format already deployed to our public WebCSD platform

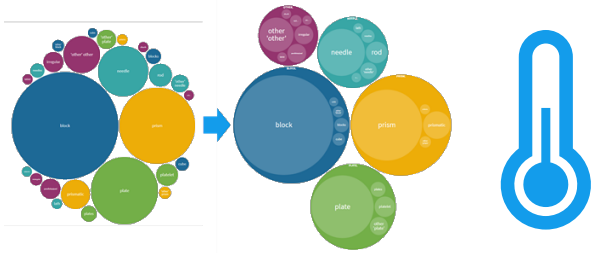
Improving existing entries

Targeted improvements allow improved integrity, consistency, discoverability and value of data

- During 2023 **over 190K** existing CSD entries have been **improved**, including improvements to:
  - Melting point, recrystallisation and bioactivity/source fields.
  - The addition of more oxidation states.
  - The labelling of radicals and polymorphs.
- New bulk editing capabilities have accelerated improvements

Enhancing Entries

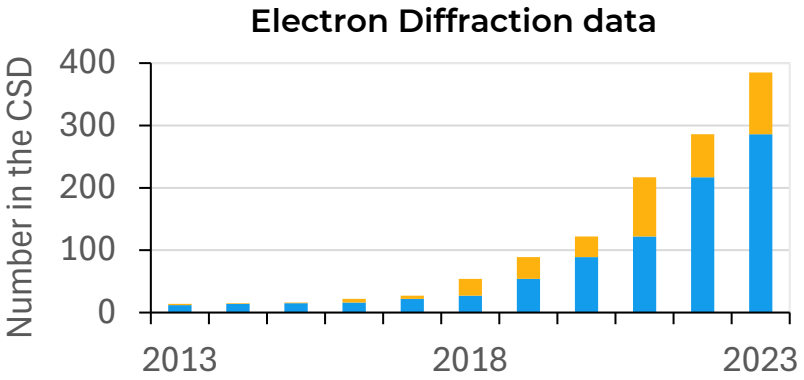


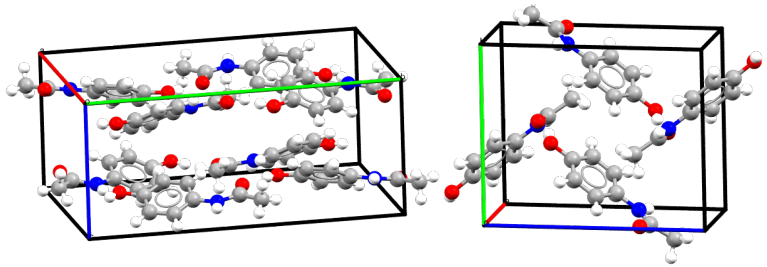


Standardising text fields e.g. habit, colour and steps to improve the AI readiness of the CSD

Keeping up with techniques

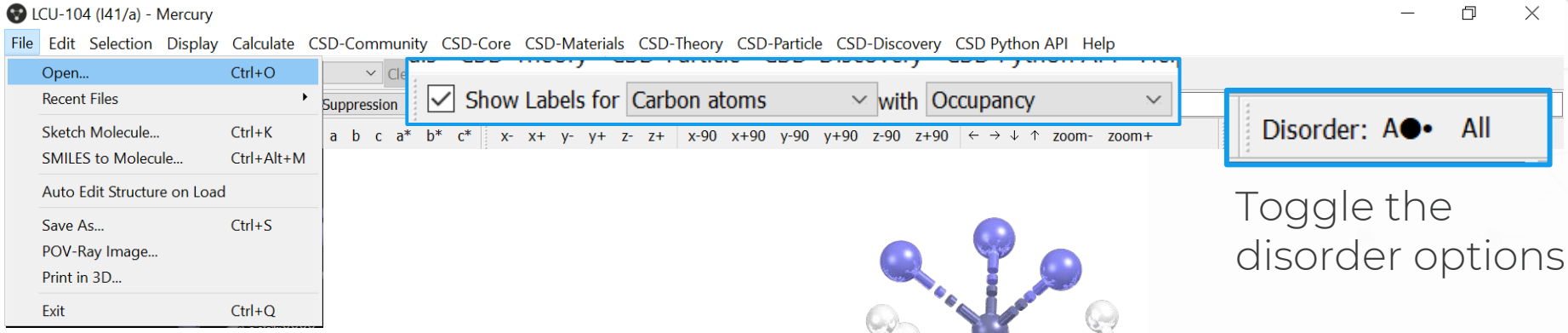
Identification method	% of CIFs
Automatic from CIF	61 %
Wavelength*	25 %
Information provided elsewhere	13 %
Only identifiable from paper	1 %



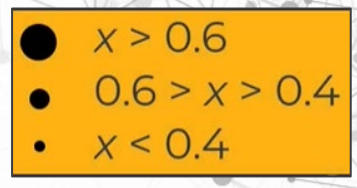


Labelling polymorphs in CSD Drug Subset

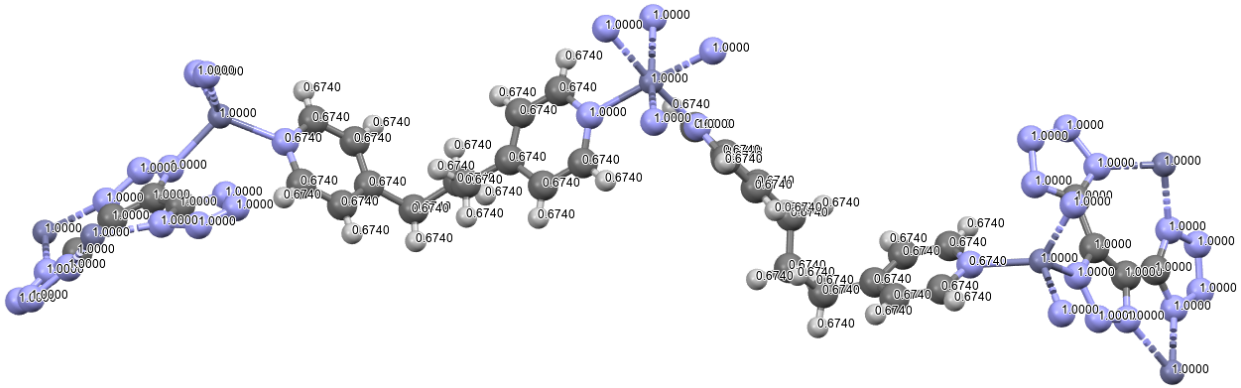
Visualising disorder

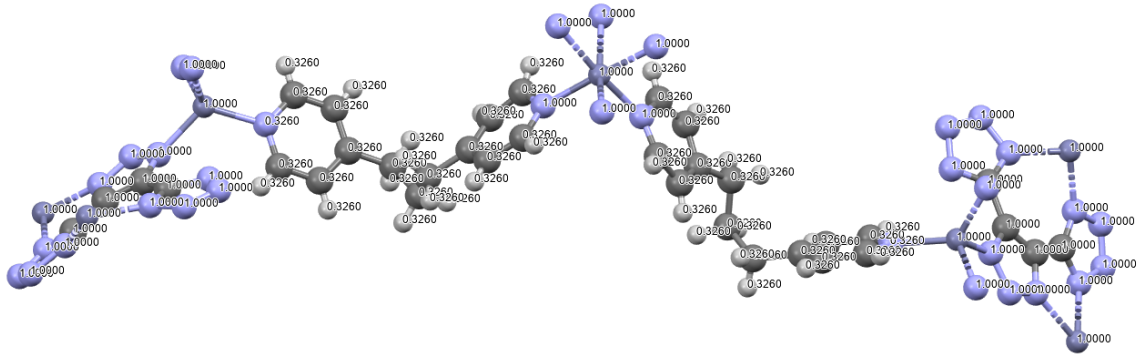


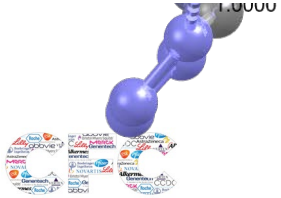
Toggle the disorder options



x = occupancy





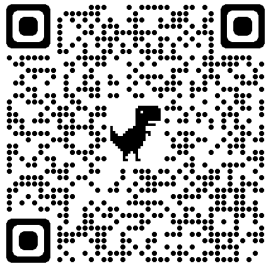


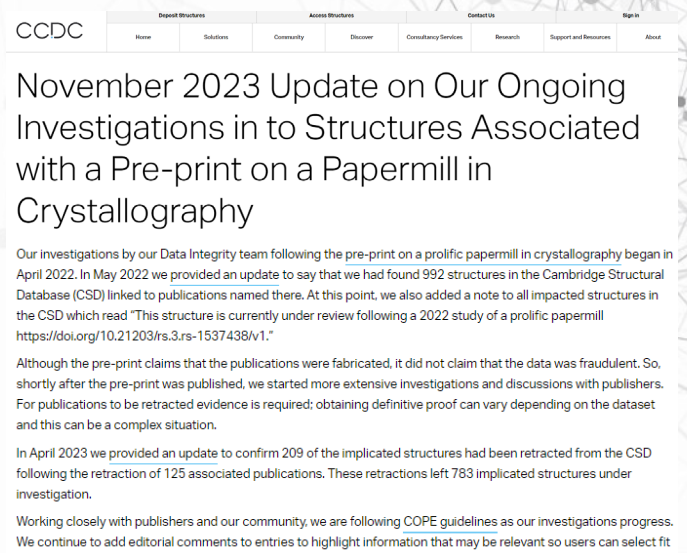
It works in combination with other functionality, including Edit, viewing and selecting hydrogen bonds, display Voids.

CCDC

Data integrity

- Online deposition process with embedded data integrity checks (checkCIF) and internal duplicate checks (automated and manual).
- Majority of data (>90%) reflects data in associated peer reviewed scientific articles. Strong relationships with major publishers, referee request service.
- Team of expert PhD level editors and scientists at CCDC, including a data integrity scientist, that support the curation of data into the CSD.
- Workflows to compare deposited datasets with the CSD through comparisons and assessments to identify some instances of fraud
- Strong connections with IUCr, individuals, some publishers, membership of COPE and use of external resources (Retraction Watch)
- Clear retraction policy and workflow





CCDC

Deposit Structures | Access Structures | Contact Us | Sign in

Home | Solutions | Community | Discover | Consultancy Services | Research | Support and Resources | About

November 2023 Update on Our Ongoing Investigations in to Structures Associated with a Pre-print on a Papermill in Crystallography

Our investigations by our Data Integrity team following the pre-print on a prolific papermill in crystallography began in April 2022. In May 2022 we provided an update to say that we had found 992 structures in the Cambridge Structural Database (CSD) linked to publications named there. At this point, we also added a note to all impacted structures in the CSD which read "This structure is currently under review following a 2022 study of a prolific papermill https://doi.org/10.21203/rs.3.rs-1537438/v1."

Although the pre-print claims that the publications were fabricated, it did not claim that the data was fraudulent. So, shortly after the pre-print was published, we started more extensive investigations and discussions with publishers. For publications to be retracted evidence is required; obtaining definitive proof can vary depending on the dataset and this can be a complex situation.

In April 2023 we provided an update to confirm 209 of the implicated structures had been retracted from the CSD following the retraction of 125 associated publications. These retractions left 783 implicated structures under investigation.

Working closely with publishers and our community, we are following COPE guidelines as our investigations progress. We continue to add editorial comments to entries to highlight information that may be relevant so users can select fit for purpose data. The CSD portfolio also enables researchers to critically evaluate the data in the context of the >1.25 million structures in the CSD. Our collaborations with publishers on investigations have also led to retractions in the scientific literature as well as the CSD. In addition, we work closely with the IUCr and other data repositories in this field.

November 2023 Status

In our last data release of 2023, scheduled for later this year, a further 152 entries implicated by the pre-print will be retracted. These retractions are already visible to users accessing the CSD via our web platforms and takes the total number of retractions related to the papermill pre-print in the CSD to 361.

We have updated the comments for 39 structures related to the papermill pre-print to say that following extensive review by the publishers and the CCDC there are currently no concerns on the data. A further 36 structures related to the papermill pre-print have been updated where our investigations have identified the entry has almost identical reflection data with another entry in the CSD with a different structural formula.

Investigations are continuing for the remaining 556 structures implicated by the pre-print. Cases where reflection data has not yet been made available can be especially challenging to assess. We will continue to work with publishers, reviewers and depositors and are grateful for how the community has come together to tackle the issues so far.

Work to broaden the automated data integrity checks and processes, conversations with publishers on how we can support their peer review processes as well as collaborations with the community remain a key priority for the CCDC.

Keep up to date with further developments here.

CCDC

Where are we now and where next?

	Candidate	Planned	In progress
Data	Capturing crystallisation conditions	New data fields available	Targeted bulk CSD improvements 2024
	Data versioning	Open licenses for deposited CIFs	Calculated semi-conductor data
	Evolving our knowledge bases	New data subsets	
	Improved metadata for PXRD, ED, etc	Visualising disorder for CSD entries	
	Alternative ways to navigate/view entries	New data format for desktop release	
	New models of data		
	Improved stereochemistry representations		
Data integrity	New integrity checks in deposition	Data integrity filters	Additional data integrity metadata
Data partners		Improved and extended links	Further publisher workflows
		CCDC visualisation in more journals	
Data standards			Community CSP data standards
			Schema for crystallisation conditions
			CoreTrustSeal re-certified

In progress

- Targeted bulk CSD improvements 2024
- Calculated semi-conductor data
- Additional data integrity metadata
- Further publisher workflows
- Community CSP data standards
- Schema for crystallisation conditions
- CoreTrustSeal re-certified

Coming up at CFC15

CCDC

Where are we now and where next?

	Candidate	Planned
Data	Capturing crystallisation conditions	New data fields available
	Data versioning	Open licenses for deposited CIFs
	Evolving our knowledge bases	New data subsets
	Improved metadata for PXRD, ED, etc	Visualising disorder for CSD entries
	Alternative ways to navigate/view entries	New data format for desktop release
	New models of data	
	Improved stereochemistry representations	
Data integrity	New integrity checks in deposition	Data integrity filters
Data partners		Improved and extended links
		CCDC visualisation in more journals

Come and chat to us and tell us your priorities for how the CSD needs to evolve

CCDC

Training courses

CSDU

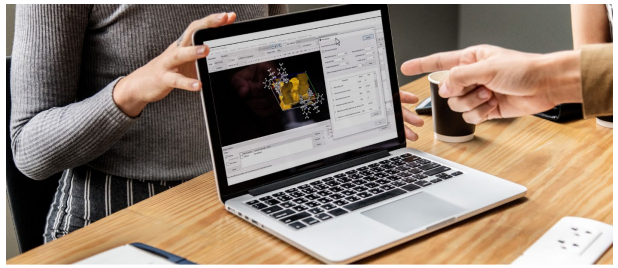
On-demand modules to learn how to use the CSD software at your own pace.

UWatch UTry UTest

Searching Mercury Python API GOLD Docking

Mogul FIMs Particle

CCDC Virtual Workshops



CCDC
FREE VIRTUAL WORKSHOPS

WHAT?

The CCDC Virtual Workshops are a series of hands-on, guided training sessions, where you learn how to use different components of CSD software. These sessions are free and open to both beginners and more experienced users of the CSD Software.

HOW?

The format is 90 minutes and *Show One, Try One, Explore More*:

- *Show One*: A guided demo of the software by the CCDC tutors.
- *Try One*: Hands-on examples for you to try with CCDC tutors on hand to help.
- *Explore More*: Learning outcomes recap, challenges, and quizzes.

APRIL-MAY 2024

16 April - 1 pm (BST)
How Crystal Structure Affects Particle Behaviour

30th April - 4 pm (BST)
Advanced Functionality for Visualization and Analysis of Structures in Mercury

14th May - 10 am (BST)
In-Depth Comparison of Polymorphic Structures Using Mercury

Registration is now open - just scan the QR code

SCAN ME

hello@ccdc.cam.ac.uk
www.ccdc.cam.ac.uk/community/events/virtual-workshops

16th April
How crystal structure affects particle behaviour

30th April
Advanced functionality for visualization and analysis of structures in Mercury

14th May
In-depth comparison of polymorphic structures using Mercury

Register now or share with your co-workers

https://www.ccdc.cam.ac.uk/community/training-and-learning/csdu-modules/

CCDC

Tributes to Olga Kennard

ACS SPRING 2024

Many Flavors of Chemistry

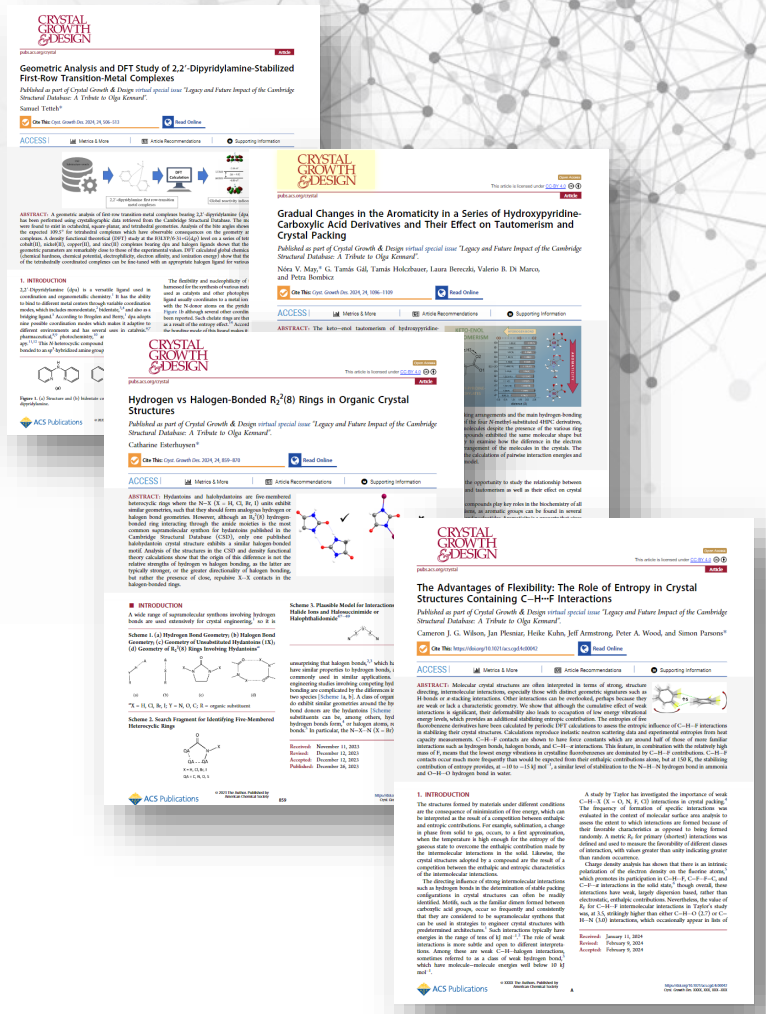
Celebrating the Life and Legacy of Dr Olga Kennard

ACS Chemistry for Life®

Sunday 17 March 2024, Spring 2024 ACS National Meeting, New Orleans

Organizers: Ian Bruno, Judith Currano, Carmen Nitsche, Suzanna Ward,

Start	Speaker	Title	
8:00	Suzanna Ward	The life of Dr Olga Kennard: A visionary leader	Remote
8:15	Stephen Burley	Protein Data Bank: From Two Epidemics to the Global Pandemic to mRNA Vaccines and Paxlovid	Remote
8:35	Rivka Isaacson	Olga Kennard as an icon of interdisciplinarity	In-person
8:55	Dirk Trauner	The chemist and the architect	In-person
9:15	Lawrence Falvello	Olga Kennard as research mentor in the mid-1970's: A DNA fragment, dipeptides, an early use of disorder tools, and cannabidiol	In-person
9:35	John Rumble	Olga Kennard and the Cambridge Crystallographic Data Centre: An Inspiration for Modern Scientific Data Science: An Outsider's View	In-person
9:55	Break		
10:05	Jeremy Sanders	Olga Kennard: The CSD and wider contributions to the community	Remote
10:20	Carolyn Brock	Approximate symmetry in organic crystal structures having $Z'>1$	In-person
10:40	Ian Bruno	From vision through data to knowledge: The insights of Dr Olga Kennard	In-person
11:00	Rajni Bhardwaj	The integration of CCDC Tools into Pharmaceutical Solid form Development	Remote
11:20	Diane Dickie	Teaching Crystallography to Chemistry Students Using the Cambridge Structural Database	In-person
11:40	Jürgen Harter	The future of structural science: Building on Olga Kennard's enduring legacy	In-person



CCDC

Thank you and questions

> “I think that the great ocean of truth is still in front of us and that we will continue to discover new aspects of this truth.”
>
> Dr Olga Kennard, Founder of the CSD

Olga Kennard - Bernal's Vision: From Data to Insight - J.D.Bernal Lecture 1995

CCDC